GRUPO EDITORIAL
DE PSICOFUNDACIÓN

Psicothema (2023) 35(3) 217-226

Psicothema

https://www.psicothema.com • ISSN 0214–9915

Psicothema

Colegio Oficial de Psicólogos del Principado de Asturias

Article

Evidence for Test Validation: A Guide for Practitioners

Stephen Sireci[1] and Isabel Benítez[2,3]

[1] University of Massachusetts Amherst, USA
[2] University of Granada, Spain
[3] Mind, Brain and Behaviour Research Center (CIMCYC), Granada, Spain

ARTICLE INFO

Received: November 16, 2022
Accepted: January 14, 2023

Keywords:
Assessment
Educational and psychological testing
Sources of validity evidence
Testing standards
Validity

ABSTRACT

Background: Validity is a core topic in educational and psychological assessment. Although there are many available resources describing the concept of validity, sources of validity evidence, and suggestions about how to obtain validity evidence; there is little guidance providing specific instructions for planning and carrying out validation studies. **Method:** In this paper we describe (a) the fundamental principles underlying test validity, (b) the process of validation, and (c) practical guidance for practitioners to plan and carry out sufficient validity research to support the use of a test for its intended purposes. **Results:** We first define validity, describe sources of validity evidence, and provide examples where each of these sources are addressed. Then, we describe a validation agenda describing steps and tasks for planning and developing validation studies. **Conclusions:** Finally, we discuss the importance of addressing validation studies from a comprehensive approach.

Evidencias Sobre la Validación de los Tests: una Guía Práctica

RESUMEN

Palabras clave:
Evaluación educativa
Evaluación psicológica
Fuentes de validez
Estándares de evaluación
Validez

Antecedentes: La validez es un tema central en la evaluación psicológica y educativa. A pesar de que la literatura disponible recoge numerosos recursos en los que se describe el concepto de validez, las fuentes de evidencia y se aportan sugerencias sobre cómo obtener evidencias de validez, apenas existen guías que proporcionen instrucciones específicas para planificar y desarrollar estudios de validación. **Método:** El presente artículo describe (a) los principios fundamentales en los que se sustenta la validez de los test, (b) el proceso de validación, y (c) una guía práctica para planificar y recoger evidencias de validez que apoyen el uso de un test para alcanzar el objetivo previsto. **Resultados:** En primer lugar, se describe el concepto de validez y las fuentes de evidencia, aportando ejemplos específicos donde se abordan cada una de ellas. A continuación, se describe una agenda de validación en la que se enumeran los pasos y tareas necesarios para planificar y completar un estudio de validación. **Conclusiones:** Finalmente, se discute la relevancia de adoptar una aproximación comprehensiva al abordar estudios de validación.

Cite as: Sireci, S. & Benítez, I. (2023). Evidence for test validation: A guide for practitioners. *Psicothema, 35*(3), 217-226. https://doi.org/10.7334/psicothema2022.477
Corresponding author: Isabel Benítez, ibenitez@ugr.es

Educational and psychological tests are widely used by educators, researchers, employers, and psychologists for a variety of important reasons such as diagnosis, treatment planning, certification, and evaluation of interventions. These uses often have important consequences such as awarding diplomas, determining placement in instructional programs, defining eligibility for services, and obtaining jobs. Tests are also increasingly used for accountability purposes where districts, schools, teachers, and even countries, are judged by how well students perform (Sireci & Greiff, 2019). Clearly, educational and psychological tests are universally valued. However, the actual value of the information provided by tests depends on the quality of the test, and the validity evidence that supports its use.

The degree to which the use and value of a test is justifiable is described by a single concept called *validity. The Standards for Educational and Psychological Testing* describe validity as "the most fundamental consideration in developing tests and evaluating tests" (American Educational Research Association [AERA] et al., 2014, p. 11). Although validity is fundamentally important, it remains mysterious to many practitioners who struggle to understand the concept and how to go about the process of test validation. As educators, psychologists, and researchers, we are compelled to ensure the assessments we administer and use have a sound, scientific basis to justify the conclusions we draw from them and the actions we take based on the information they provide. Therefore, a comprehensive understanding of what validity is and how to evaluate validity evidence is crucial for effective research and practice in the social sciences. In this article, we describe the concept of validity, provide practical guidance and examples for evaluating the validity of a test for a particular purpose, and guidance for conducting test validation research.

Defining Validity

The AERA et al. (2014, 2018) *Standards* define validity as "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). This definition emphasizes validity is not something inherent within a test, but rather refers to the process of using a test for a specific purpose. Thus, a particular test may be valid (i.e., supported by theory and evidence) for one purpose, but invalid for another.

We use and follow the AERA et al. (2014) definition of validity because it is based on over 60 years of collaborative work across three professional associations (Sireci, 2020) and so represents a comprehensive consensus. However, as Newton and Shaw (2013) pointed out, there are debates and contentions in the validity literature, with some arguing test use is less relevant to validity; and what really matters is evidence the test measures what it intends to measure (see Sireci 2016a, 2016b). Although such evidence is important to *support* the use of a test for a particular purpose, it is not *sufficient* for doing so. As the *Standards* (AERA et al., 2014) and others have pointed out (e.g., Kane, 2006; Messick, 1989; Mislevy, 2019; Shepard, 1993), it is the actions made on the basis of test scores that affect *people,* and so it is these actions that must be justified by validity evidence. The more comprehensive view of validity promulgated by the Standards is consistent with the ethical principles of psychologists and educators to first, "do no harm" (American Psychological Association [APA], 2010, p. 3), and so is the perspective we emphasize here.

Sources of Validity Evidence

The AERA et al. (2014) *Standards* specify five sources of validity evidence "that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). These five sources are validity evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) testing consequences. In this section, we provide brief descriptions of these sources of evidence. As described subsequently, a proper validation of test use requires synthesizing all accumulated evidence into a "validity argument" (Kane, 1992, 2013) that can be evaluated to judge whether use of the test is sufficiently justified.

Understanding "The Construct"

Before describing the five sources of validity evidence, we first define the term *construct* because it permeates all test development and validation. The term "construct" refers to the knowledge and skill domain measured on an educational test, or to another personal attribute measured by a psychological test. Cronbach and Meehl (1955) introduced this term to refer to "some postulated attribute of people, assumed to be reflected in test performance" (p. 283). Thus, the construct is what test developers aim to measure, and the conceptual framework within which test scores are interpreted.

Validity Evidence Based on Test Content

Validity evidence based on test content evaluates the extent to which the content of the test represents the intended construct and is consistent with the testing purpose (Sireci & Faulkner-Bond, 2014). Thus, acquiring validity evidence based on test content involves gathering information about both the intended construct and the test, and collecting evidence about the overlap between them. Traditionally, this is addressed by asking experts to judge the adequacy of the test content using procedures that capture the intended information in a systematic and standardized way (Beck, 2020; Martone & Sireci, 2009; Sireci, 1998). For example, experts may evaluate the degree to which items on an employment test are relevant to the tasks employees conduct on the job.

Example of Validity Evidence Based on Test Content

Table 1 presents an example of a rating form given to subject matter experts to evaluate the validity of the content for an educational test. This rating form was used to evaluate the degree to which the items on reading achievement tests for adults (the Massachusetts Adult Proficiency Tests, or MAPT, see Zenisky et al., 2018) were measuring the curriculum objectives they intended to measure. As implied in the third column of Table 1, the instructions to the expert reading teachers were simple—review each item, consider the objective it was designed to measure, and rate how well the item measures it. The illustration in Table 1 only lists two items on the rating sheet for illustrative purposes. A summary of the results from the actual study, which involved rating 1,370 items, is presented in Table 2.

As indicated in Table 2, using a criterion of a median rating of 4.0 on the six-point scale, 73% of the items were considered to adequate measuring their intended objective. The other items were classified in need of revision or elimination from the item bank (for details see Zenisky et al., 2018). This process improved the content relevance of the assessment and provided valuable validity evidence based on test content.

Table 1
Example of Subject Matter Expert Content Validity Rating Form

Objective measured	Description	How well does the item measure its objective? (Circle One) 1 = Not at All; 6 = Very Well	Comment(s)
CCRSAE-4.B.1	Determine the meaning of general academic and domain-specific words and phrases in a text relevant to a topic or subject area. (RI.3.4)	1 2 3 4 5 6	
CCRSAE-5.B.2	Use text features and search tools (e.g., key words, sidebars, hyperlinks) to locate information relevant to a given topic efficiently. (RI.3.5)	1 2 3 4 5 6	

Studies to gather validity evidence based on test content can be much more comprehensive. For example, alignment studies require subject matter experts to make several judgements such as how well items align to their objectives, the sufficiency with which the items represent the intended objectives, the degree to which the intended cognitive levels are measured, and the degree to which the difficulty of the items is appropriate for the testing purpose. An example of how the results of an alignment study provide validity evidence based on test content is presented in Table 3 (from Sireci et al., 2018). This study used seven alignment criteria such as whether items had median ratings of "good" or better with respect to how well they measured the test's content areas (column 3) and whether the items judged to represent their intended objectives sufficiently represented the intended test specifications (within a 5% tolerance limit; column 9). There are a variety of methods and criteria for evaluating the results of alignment studies (see Bhola et al., 2003; Martone & Sireci, 2009; Reynolds & Moncaleano, 2021).

Table 2
Frequency of Median Item Ratings: Reading Content Validity Study

Median Rating Category	Number of Items	Percent of Items
1-1.9	230	16%
2-2.9	55	4%
3-3.9	104	7%
4-4.9	81	6%
5-5.9	217	15%
6	751	52%
Total Medians ≥ 4	1,049	73%

Validity Evidence Based on Response Processes

Validity evidence based on response processes pertains to the overlap between the intended construct and participants' responses (AERA et al., 2014). The thought processes used by participants while responding to items show whether their interpretations reflect the intended construct. Therefore, gathering this type of validity evidence involves the identification of the elements that cause responses. Qualitative procedures have been extensively used to learn how participants understand the question, recover the information, make a judgement and report their answer. Among them, interviews (i.e., Noble et al., 2014) and, more specifically, cognitive interviewing (i.e., Cavalcanti et al., 2020; Padilla & Benítez, 2014) have demonstrated utility. However, other procedures, such as processing time (Engelhardt & Goldhammer, 2019), behavior coding (Rushton et al., 2015), and eye-movement indices (Lee & Winke, 2018), have been applied to analyze the overlap between the participants' response processes and the intended construct.

Table 3
Summary of Alignment Evaluation

Subject	Grade	Criterion						
		Categorical Concurrence				Range of Knowledge	DOK Consistency	Domain Representation
		> 80% items median > Good	> 80% items measuring *Standard*	> 80% items aligned with *Area*	> 6 items each *Standard*	> 50% Areas Measured by > 1 item	> 50% items at DOK or higher	Represent test specifications +/-5%
Math	3	Met	Met	Met	Met	Met	Met	Met
	4	Met	Met	Met	Met	Met	Met	Met
	5	Met	Met	Met	Met	Met	Met	Met
	6	Met	Met	Not Met	Met	Met	Met	4 of 5 Met
	7	Met	Met	Met	3 of 4 Met	Met	Met	Met
	8	Met	Met	Met	Met	Met	Met	Met
	11	Met	Met	Met	Met	3 of 5 Met	Met	Met
Science	4	Met	Met	Met	Met	Met	Met	Met
	8	Met	Met	Met	Met	Met	Met	2 of 3 Met
	11	Met	Met	Met	Met	3 of 4 Met	Met	Met

Example of Validity Evidence Based on Response Processes

Table 4 shows a section from an interview protocol that collected information about the processes used by participants while responding to a scale measuring quality of life; specifically, when responding to an item asking about the importance of family in their life (see Benítez et al., 2022). Participants first responded to a set of questions where they rated the importance of different issues in their lives. Then, the interviewer explored how they reached the answer and the strategies for deciding their response. The protocol in Table 4 illustrates the probes formulated to understand the response processes when answering the item about importance of the family.

Table 4
Example of Interview Protocol for Capturing Response Processes

Introduction to the section	First, I asked you some questions about how important these aspects were in your life, for example work, family and acquaintances.
General probe	How did you respond these questions in general?
Specific probes	One of the questions asked about "family." Which people did you think about when responding? How did you select the alternative that best reflected your situation?

As illustrated in Table 4, first the interviewer contextualized the question to be inquired. Then, the formulation of a general question was made to understand the general process to respond to the different statements. In this probe, participants described general strategies such as how they used the response options. For instance, some participants explained how they made "relative comparisons" among statements rating first the statement "work" (placed in the first position in the scale) and then compared the rest, deciding whether "family" (the following statement) was more or less important than work. Thus, the general probe gave a global perspective about response processes related to the scale, and the specific probes provided details about potential causes of the unexpected differences between groups, such as the different use of the term "family" in the two countries. While the Spanish term for family (familia) includes close and extended family for Spaniards, the Dutch term (gezin) just refers to nuclear family. Dutch participants habitually mentioned to their household members in arguments as "I thought of my family and home.... My brother, my sister, my mother…"; whereas Spanish participants considered other family members saying sentences as "I thought of my wife, my daughter, my father and other relatives of my father." Information collected during the interviews provided evidence for understanding why the participants' responses were not equivalent and, therefore, incomparable (see Benítez et al., 2022 for further discussion).

The previous example shows how analysis of participants' response processes provides evidence of validity, as it informs us about the elements participants considered when responding. When the evidence illustrates sources of problematic or unexpected interpretations, the assessment needs to be improved. When searching for response processes to support test use, the procedure will focus on determining situations in which participants interpreted the item as expected, and situations in which they did not, as well as arguments for both cases. These alternative scenarios are valuable for improving the assessment.

Validity Evidence Based on Internal Structure

Validity evidence based on internal structure investigates how the relationships among items and the dimensions underlying the test support the proposed interpretation of the scores (AERA et al., 2014). Such evidence evaluates the connections between the test components to ensure the constructs are represented by the scores (and subscores). Therefore, providing validity evidence based on internal structure requires collecting responses from participants to analyze how these responses reflect the intended test structure. Rios and Wells (2014) described the main methods for extracting validity evidence about the internal structure of an assessment as focusing on evaluating dimensionality, measurement invariance, and reliability. Among them, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are two common procedures for analyzing the dimensional structure of tests (see Ferrando et al., 2022 for details). Special attention has also been paid to the presence of DIF, which AERA et al. (2014) classify as validity evidence based on internal structure, but can be also interpreted in terms of other sources of validity evidence (Gómez-Benito et al., 2018).

Examples of Validity Evidence Based on Internal Structure

Results gathered through dimensionality analysis provide validity evidence when they reveal to what extent responses reflect the theoretical composition of the construct. For instance, Lafuente et al. (2022) used CFA to understand the dimensionality of the construct "computational thinking." Due to the extensive literature defining the construct and the diversity of theoretical approaches describing its dimensions, they used CFA to evaluate three of the most defended models. Although the three structures reached adequate fit values, the item loadings pointed to a unidimensional model measuring "algorithmic thinking." Thus, the evidence supported using the test of computational thinking for measuring a single dimension, rather than assessing the originally hypothesized multidimensional construct.

Validity evidence based on internal structure is especially useful for studying cultural or group impacts on test score interpretations. DIF studies are particularly helpful for evaluating whether items are consistently measuring the same construct across different groups of test takers. DIF is especially important nowadays, due to the fact that international and comparative studies are crucial in educational and psychological testing. Identification of DIF is a statistical procedure, but interpreting the results requires additional research. One illustration of a comprehensive validity analysis based on DIF is Benítez et al. (2016) who analyzed DIF in seven scales from the PISA Student Questionnaire. They flagged several items for DIF across students from Spain and the United States. Additional insight came from asking experts to evaluate the U.S. and Spanish versions of the items to identify elements causing the DIF. The information provided by the experts identified problematic issues and argued against the use of the scales for comparing these groups. For instance, experts pointed out differences between the English and the Spanish versions of the item "I like reading about science;" the English version asked about the activity of reading in general, and the Spanish version about reading books of science (see Benítez et al., 2016 for additional examples).

Evidence based on internal structure is frequently used in validation studies as it can be obtained by analyzing the responses collected from participants during pilot or operational testing. Internal structure validity studies should focus not only on replicating the theoretical structure of the test, but also on understanding the organization of the elements in the participants' responses.

Validity Evidence Based on Relations to Other Variables

Validity evidence based on relations to other variables refers to studies that involve test scores analyzed together with other variables in correlational, experimental, or quasi-experimental designs. This form of validity evidence was historically referred to as "criterion-related validity" (Sireci, 2020), which was often partitioned into concurrent validity (test and criterion data are gathered at similar points in time) or predictive validity (criterion data are gathered well after examinees complete a test). However, as Messick (1989) pointed out, this type of validity evidence could include analysis of group differences in test scores, particularly when the groups are manipulated in experimental studies (e.g., pretest-posttest, or randomly equivalent group designs).

There are many examples of studies that gather validity evidence using external criteria. One example can be taken from the MAPT Reading and Math tests, where students' MAPT scores were correlated with their scores on high school equivalency tests called the HiSET (Zenisky et al., 2018). The MAPT is not designed to predict HiSET scores, but both tests are measuring similar constructs—math and reading. The correlations among the different sections of each test, are presented in Table 5. The study involved 178 students who had taken both tests, and the within-subject correlations (Reading/Reading and Math/Math) were higher (.73 and .67) than the across subject correlations (Reading/Math, which ranged from .45 to .50). This pattern of correlations supports the convergent (relatively higher correlations across measures of similar constructs) and discriminant (relatively lower correlations among measures of dissimilar constructs) validity (Campbell & Fiske, 1959).

Validity evidence based on relations to other variables is strongest when the study design focuses on the theory underlying the test (i.e., the construct intended to be measured) and how the test scores are expected to relate to the external variables included in the analysis (Messick, 1989).

Table 5
MAPT and HiSET Reading and Mathematics Score Correlations (n=178)

		Mean	SD	MAPT-CCR		HiSET	
				Reading	Math	Reading	Math
MAPT-CCR	Reading	567.3	76.58	-			
	Math	574.2	77.49	.50*	-		
HiSet	Reading	11.3	4.34	.73*	.45*	-	
	Math	10.5	4.36	.47*	.67*	.47*	-

Note: From Zenisky et al. (2018).
*p<.001.

Validity Evidence Based on Testing Consequences

Validity evidence based on testing consequences aims to identify and evaluate consequences derived from the use of test scores and their interpretations (Messick, 1989). Both positive and negative consequences, intended and unintended, are included in the evaluation. The AERA et al. (2014) *Standards* state, "Some consequences of test use follow directly from the interpretation of test scores for uses intended by the test developers... A fundamental purpose of validation is to indicate whether these specific benefits are likely to be realized" (p. 19).

Gathering validity evidence based on testing consequences involves analyzing the links between the test performance, the conclusions derived from the interpretations of the test scores, and the decisions made based on these interpretations. Qualitative procedures such as cognitive interviews and focus groups are often used as well as surveys of intended stakeholders (e.g., students, teachers, etc., Lane, 2014). Quantitative procedures, such as the evaluation of adverse impact (Sireci & Geisinger, 1998) and the use of Pareto curves in weighting test scores in selection decisions (De Corte et al., 2007; Newman et al., 2022) are also helpful for evaluating and addressing negative consequences (Dumas et al., 2022).

In Table 6 we present a listing of the types of studies that can be done to evaluate testing consequences, along with citations illustrating how to conduct these studies, interpret their results, and relate the results to the intended testing purposes. Analysis of adverse (disparate) impact involves looking at the percentages of examinees who pass the test or earn a job or some other benefit based on their test performance. According to the United States Equal Employment Opportunity Commission (EEOC) *Guidelines,* disparate impact exists when the certification rate for members of a protected group is less than 80% of the certification rate for the group with the highest rate of certification (U.S. Equal Employment Opportunity Commission, 2010). This "four-fifths" rule is used to signal when the use of a test can be considered to have negative consequences at a level high enough to challenge its use test in court.

The use of surveys or focus groups to evaluate testing consequences can take many forms such as asking students whether they are stressed when taking tests or discouraged by test results. Surveys of teachers can analyze teaching practices perhaps before and after tests are mandated, to see the degree to which teaching practices are affected by the test. Such surveys can evaluate whether the changes in teaching practices are positive (as would be intended) or negative.

Figure 1 provides general examples of procedures to gather the five sources of validity evidence.

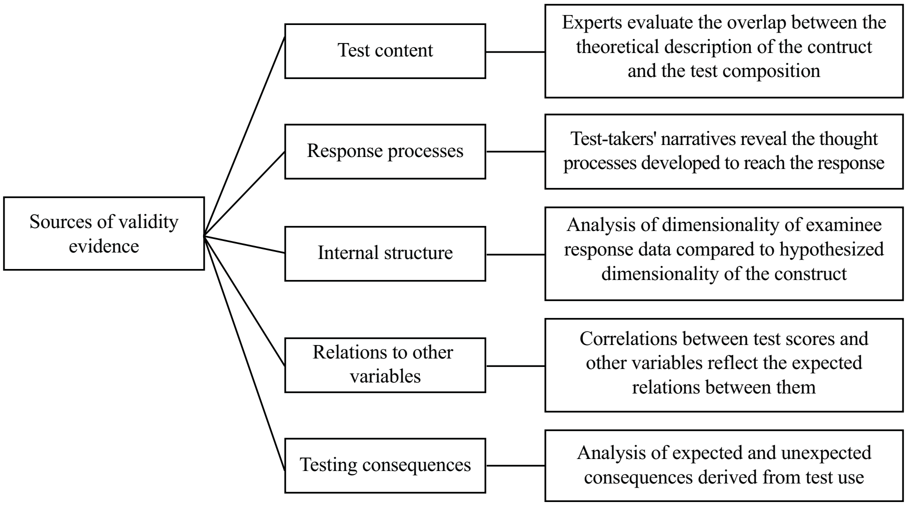
Conducting Validation Research

In the previous sections we described sources of validity evidence and gave examples of applied studies for each source. However, a single validity study is not sufficient to defend the use of a test for an indented purpose. Cizek et al. (2008) pointed out that the absence of a specific plan for conducting a validation study could lead to the risk of providing only the most easily accessible sources of validity evidence. This warning underscores the importance of distinguishing between available information that could be "used" as validity evidence and what is actually needed to provide accurate information about an instrument and sufficient justification for its applied use. In the next section, we provide a guide for planning validation studies where the starting point is reflecting on the information needed to support use of the test results for their intended purposes, and for summarizing the results across validation studies to develop a validity argument (Kane, 2013). We describe the steps to be followed for a comprehensive validation.

Table 6
Examples of Studies of Validity Evidence Based on Testing Consequences

Testing Purpose	Type of Study	Research Questions	Example References
Evaluate students' mastery of curriculum	Curriculum surveys	Are teachers teaching the intended curriculum? Is teaching to the test narrowing the curriculum	Atchison et al. (2022)
	Longitudinal analysis of student achievement	Are the intended improvements in achievement being realized?	Dee & Jacob (2011); Whitney & Candelaria (2017)
Admissions Test	Analysis of adverse impact	Do admissions tests result in disproportionate admissions for applicants from historically minoritized groups?	Sinha et al. (2011)
	Analysis of college admissions data	Does requiring all high school students to take a college admissions test result in more students going to college? More diversity within colleges?	Marchant & Paulson (2005)
Certification Test	Analysis of adverse impact	Do tests result in disproportionate passing rates for applicants from historically minoritized groups?	Morris & Dunleavy (2016)
	Audit of test development procedures	Is the content of the test from a dominant culture perspective that inhibits the performance of minority candidates?	Randall (2021)
Accountability	Student surveys	Are students stressed out when taking the tests? Are certain types of students discouraged by the results reported?	Segool et al. (2013)
	Teacher surveys	Are teachers teaching to the test at the expense of a broader curriculum? Do teachers find the test results helpful?	Zenisky et al. (2018)
	Longitudinal analysis of student achievement	Has student achievement improved over time?	Irwin et al. (2022)
Diagnosis of disability	Comparison of diagnostic assessments	Do different tests of reading disabilities lead to different classifications of students?	Keenan & Meenan (2014)

Figure 1
General Examples of Sources of Validity Evidence


Sources of validity evidence
Test content
Experts evaluate the overlap between the theoretical description of the contruct and the test composition
Response processes
Test-takers' narratives reveal the thought processes developed to reach the response
Internal structure
Analysis of dimensionality of examinee response data compared to hypothesized dimensionality of the construct
Relations to other variables
Correlations between test scores and other variables reflect the expected relations between them
Testing consequences
Analysis of expected and unexpected consequences derived from test use

Steps for Planning a Validation Agenda

Step 1: Clearly Define the Intended Purposes of the Test

A key step in test validation is to adequately articulate the purposes of the test. This articulation determines how scores will be interpreted and used, and sets the goals of what is to be validated (AERA et al., 2014). In educational assessments the testing purposes are frequently connected to the expected learning of students or to established requirements for being admitted in a specific program. In these cases, the purposes must be defined in terms of abilities or skills. For example, the purposes of the MAPT tests, described earlier, are:

> ...to measure [adult education students'] knowledge and skills in mathematics and reading so their progress in meeting educational goals can be evaluated. The MAPT is designed to measure learners' educational gains for the purposes of state monitoring and [Federal] accountability... MAPT scores and score gains can be aggregated to provide meaningful summative measures of program effectiveness. (Zenisky et al., 2018, p. 10)

For some tests, such as those measuring non-cognitive variables, the purpose of the test could refer classification of examinees. For instance, in the Spanish version of the *Intellectual Humility Scale* (Luesia et al., 2021) the purpose is to measure participants' intellectual humility for analyzing the relationship of intellectual humility with other variables (e.g., academic achievement) and to compare participants.

Step 2: Identify Potential Misuses and Potential Negative Consequences

It is also important to ensure use of the test does not cause unintended negative consequences. Thus, it is important to conduct research to anticipate any potential negative consequences and organize the testing processes to mitigate against them. Identifying potential negative consequences may simply involve being aware of common criticisms of testing programs and surveying invested stakeholders. For example, if the test is a licensure test for physicians, surveys of currently licensed doctors, candidates for licensure, hospital staff, and medical school educators may identify potential negative consequences such as overly strict testing conditions, prohibitive costs to take the exam that leads to exclusion of less privileged candidates, and so forth. Narrowing the curriculum is also a criticism of educational tests, and that potential negative consequence can be studied by surveying teachers, through classroom observation, and analysis of lesson plans (Lane, 2014).

A common criticism across high-stakes tests, such as admissions tests, employment tests, and credentialing exams; is disparate (adverse) impact across historically privileged and historically marginalized groups. For that reason, analysis of potential adverse impact should be planned from the earliest stages of a testing program.

Step 3: Determine the Most Appropriate Sources of Validity Evidence

The next step is to identify the evidence needed to support each intended test use and evaluate the potential negative consequences. Each source of validity evidence needed to support each test use or evaluate negative consequences should be identified. An example of this identification is presented in Table 7. The research purposes associated with each source of validity evidence are listed as are the procedures and data needed to conduct the validity studies related to each source. Considering the need and value of each source of validity helps decide how to approach validation and gather the most appropriate evidence for a given validity argument.

Step 4: Conducting the Validity Studies

After the studies have been planned, the data should be gathered and analyzed according to the plan. For most sources of evidence data collection and analysis will occur after the test has been administered, or at least piloted. However, for validity evidence based on test content, subject matter experts' appraisal can be gathered during various stages of test development.

Step 5: Integrating the Validity Evidence to Create a Validity Argument

To reach a comprehensive picture of test functioning requires assessing the consistency of information coming from the various sources. As the *Standards* claim, "A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses" (AERA et al., 2014, p. 21). Therefore, "Validation can be viewed as a process of constructing and evaluating arguments for and against the intended interpretation of test scores and their relevance to the proposed use" (AERA et al., 2014, p. 11).

Table 7
Validity Research Framework

	Source of Validity Evidence				
	Test content	**Internal structure**	**Rel. to other variables**	**Response processes**	**Testing conseq.**
Research Purpose	Evaluate the overlap between the construct definition and the test content	Evaluate whether intended test structure is recovered via analysis of responses.	Discover the relationships between the test and other measures	Identify incongruences between the intended construct and participants' responses	Confirm intended consequences and no negative consequences
Procedures	Experts evaluate and rate items and test components	Assessment of dimensionality and measurement invariance; estimates of precision	Correlation; regression; group comparisons	Cognitive interviews; eye-movement; log data analysis	Interviews; focus groups; surveys; instructional artifacts
Evidence used to…	Confirm construct representation and content is consistent with testing purposes	Confirm test structure, understand dimensionality and consistency of dimensionality across subgroups; confirm fit of scaling model.	Explore convergent & discriminant validity; test validity hypotheses (e.g., differential predictive validity)	Confirm intended response processes are elicited; evaluate participants' interpretations of items	Evaluate utility and fairness of the decisions made from test scores
What information do we need?	Subject matter expert ratings	Participants responses to items; theoretical structure of the test	Test scores and criterion scores for participants	Information about mental processes used by participants when responding to items	Adverse impact data; survey/ interview results; curricular, dropout, & other trend data
Output	Experts' ratings of how well items and other components represent construct and testing purposes	Quantitative and visual characterization of data structure; model-data fit indices; reliability estimates; test information	Statistical tests of direction & strength of connections between variables & groups); descriptive statistics	Participants' verbal responses, physical responses, log data	Adverse impact, survey, interview, trend, and other results
Interpretations, actions, and decisions	Eliminate problem items, write new items for under-represented areas, confirm construct representation	Evaluation of fit between observed and hypothesized structure; proportion of estimated true score variance	Are expected hypotheses & relationships supported by the results?	Do participants use the expected thought processes and interpret items as intended?	Are the intended testing purposes being realized? Are there negative consequences?

Synthesizing the different sources of validity evidence into a validity argument begins with restating the testing purposes, and then illustrating how each source of evidence provides information to support use of the test scores for their intended purposes. For example, an elementary school math test tied to a national curriculum could provide validity evidence based on test content to show experts agreed with the content areas measured on the test and confirmed the items adequately represents the national curriculum; validity evidence based on response processes could confirm the intended cognitive skills were being measured; validity evidence based on internal structure could confirm the unidimensional IRT scaling model demonstrated sufficient fit to students' responses to the items and estimates of score reliability were high; validity evidence based on relations to other variables could show the test scores correlated strongly with students' math grades and less so with their reading grades; and validity evidence based on testing consequences could show the test scores helped math teachers target their instruction to best help each individual student.

Summarizing and reporting validity results can be organized using the same framework used to develop the validity agenda—the five sources of validity evidence. The validity studies can be summarized under each evidence category, with each summary describing how the evidence supports (or refutes, or qualifies) use of the test for its intended purposes. Examples of organizing the validity argument in this way can be found in Georgia Department of Education and Data Recognition Corporation (2019) and Zenisky et al. (2018).

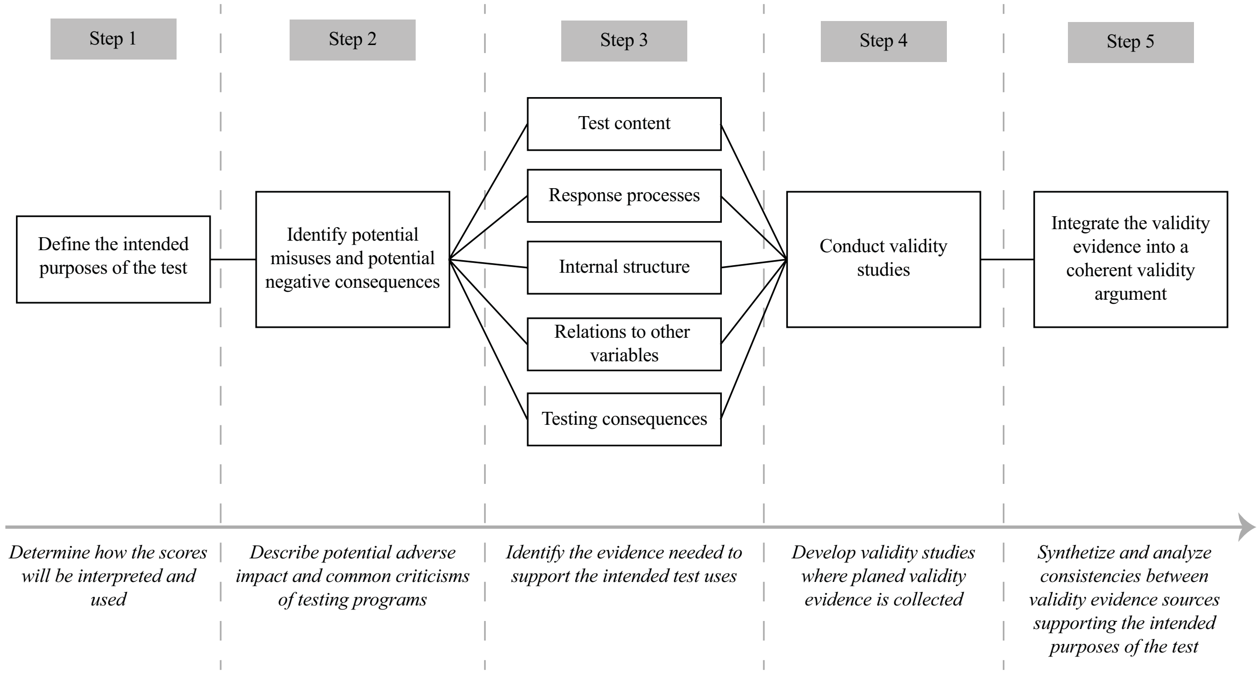
It should be noted that a single source of validity evidence is unlikely to provide a compelling validity argument, and a strong validity argument could be made without drawing on all five sources of evidence. This latter outcome is particularly likely for newer testing programs where there has been less time to conduct validity studies, particularly those that involve external variables and testing consequences. It should also be noted that a comprehensive validity argument, no matter how carefully constructed, and regardless of the weight of evidence, will never be perfect, and can never "prove" the validity of the use of a test for a particular purpose in an absolute sense. Instead, as the AERA (2014) *Standards* point out, "…at some point validation evidence allows for a summary judgment of the intended interpretation that is well supported and defensible" (p. 22). Although an absolute judgment is not possible, the validity argument should make most test users and stakeholders conclude use of the test is justifiable for its intended purposes. A compelling validity argument will draw from the five sources of validity evidence to "include evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all test takers, as appropriate to the test interpretation in question" (AERA et al., 2014, p. 22).

Figure 2 illustrates the steps for planning a validation agenda and summarizes the key activity in each of them.

Discussion

In this article, we defined validity and gave advice and examples on how to gather different sources of validity evidence, and how to synthesize the evidence into a validity argument. Our proposal is based on previous contributions, such as the argument-based approach proposed by Kane (1992, 2006, 2013) and examples in the literature where the searching of validity evidence was guided by testing purposes and criticisms.

Figure 2
Steps for Developing a Comprehensive Validation Study


Step 1
Step 2
Step 3
Step 4
Step 5
Test content
Response processes
Internal structure
Relations to other variables
Testing consequences
Define the intended purposes of the test
Identify potential misuses and potential negative consequences
Conduct validity studies
Integrate the validity evidence into a coherent validity argument
Determine how the scores will be interpreted and used
Describe potential adverse impact and common criticisms of testing programs
Identify the evidence needed to support the intended test uses
Develop validity studies where planed validity evidence is collected
Synthetize and analyze consistencies between validity evidence sources supporting the intended purposes of the test

Our proposal for synthesizing validity evidence for validation uses the AERA et al. (2014) sources of validity evidence as a guiding framework, with each source of evidence targeted to evaluating a specific test use or negative consequence. Thus, the validation framework we proposed addresses both intended testing purposes and anticipated limitations to help plan meaningful validation studies beyond merely conducting analyses based on available data. Our proposed framework places the intended use of test score interpretations in the center, because tests are developed for specific, intended uses, and the test scores must be properly interpreted to justify appropriate test use. Therefore, validation studies can be easily planned and designed when tests are developed following a systematic framework (see Muñiz & Fonseca-Pedrero, 2019, for details).

As many previous psychometricians argued, validating the use of a test for its intended purposes requires evidence confirming the test measures what it claims to measure, evidence its intended purposes are being realized, and evidence unintended negative effects are not occurring (Cronbach, 1971; Messick, 1989; Russell, 2022; Shepard, 1993). The AERA et al. (2014) *Standards*, as well as other professional guidelines (e.g., International Test Commission & Association of Test Publishers, 2022) also emphasize the importance of evaluating both intended and unintended consequences of testing in test validation. Regardless of whether practitioners follow the exact steps we proposed and illustrated, defensible test validation requires: (a) clear identification of the intended interpretations and uses of test scores, (b) validity evidence to support those interpretations and uses, (c) validity evidence to ensure the absence or minimization of unintended negative consequences, and (d) synthesis of the various sources of validity evidence into a coherent rationale (argument) that confirms and defends the use of the test. Meeting these requirements will lead to more sound and valid testing practices for all educational and psychological assessments.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association. https://www.apa.org/science/programs/testing/standards.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* [Standards for educational and psychological testing]. https://www.testingstandards.net/uploads/7/6/6/4/76643089/9780935302745_web.pdf

American Psychological Association (2010). *Ethical principles of psychologists and code of conduct*. Author. https://doi.org/10.1037/amp0000102

Atchison, D., Garet, M. S., Smith, T. M., & Song, M. (2022). The validity of measures of instructional alignment with state standards based on surveys of enacted curriculum. *AERA Open, 8*, 1-17. https://doi.org/10.1177/23328584221098761.

Beck, K. (2020). Ensuring content validity of psychological and educational tests--the role of experts. *Frontline Learning Research, 8*(6), 1-37. https://doi.org/10.14786/flr.v8i6.517

Benítez, I., Van de Vijver, F., & Padilla, J. L. (2022). A mixed methods approach to the analysis of bias in cross-cultural studies. *Sociological Methods & Research, 51*(1), 237-270. https://doi.org/10.1177/0049124119852390

Benítez, I., Padilla, J.L., Hidalgo Montesinos, M. D., & Sireci, S. G. (2016). Using mixed methods to interpret differential item functioning. *Applied Measurement in Education, 29*(1), 1-16. https://doi.org/10.1080/08957347.2015.1102915

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice, 22*(3), 21-29.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

Cavalcanti, R. V. A., Junior, H. V. M., de Araújo Pernambuco, L., & de Lima, K. C. (2020). Screening for masticatory disorders in older adults (SMDOA): An epidemiological tool. *Journal of Prosthodontic Research, 64*(3), 243-249. https://doi.org/10.1016/j.jpor.2019.07.011

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Cronbach, L. J. (1971). *Test Validation. In R.L. Thorndike (Ed.) Educational measurement* (2nd ed., pp. 443-507). American Council on Education.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

De Corte, W., Lievens, F., & Sackett, P. R. (2007). Combining predictors to achieve optimal trade-offs between selection quality and adverse impact. *Journal of Applied Psychology, 92*(5), 1380. http://doi.org/10.1037/0021-9010.92.5.1380

Dee, T. S., & Jacob. B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*, 418-446. http://doi.org/10.1002/pam.20586

Dumas, D., Dong, Y., & McNeish, D. (2022). How Fair is my Test? A Ratio Coefficient to Help Represent Consequential Validity. *European Journal of Psychological Assessment, 0*(0), 1-25. https://doi.org/10.1027/1015-5759/a000724

Engelhardt, L., & Goldhammer, F. (2019). Validating test score interpretations using time information. *Frontiers in Psychology, 10*, Article 1131. https://doi.org/10.3389/fpsyg.2019.01131

Ferrando, P. J., Lorenzo Seva, U., Hernández Dorado, A., & Muñiz, J. (2022). Decalogue for the factor analysis of test items. *Psicothema, 34*(1), 7-17. https://doi.org/10.7334/psicothema2021.456

Georgia Department of Education and Data Recognition Corporation. (2019). *Georgia Milestones Assessment System 2019 operational technical report. Georgia* Department of Education.

Gómez-Benito, J., Sireci, S., Padilla, J.L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104-109. https://doi.org/10.7334/psicothema2017.183.

International Test Commission & Association of Test Publishers (2022). Guidelines for technology-based assessment. *International Test Commission*. https://www.intestcom.org/page/28.

Irwin, V., De La Rosa, J., Wang, K., Hein, S., Zhang, J., Burr, R., Roberts, A., Barmer, A., Bullock Mann, F., Dilig, R., & Parker, S. (2022). *Report on the Condition of Education 2022* (NCES 2022-144). National Center for Education Statistics. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2022144

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535.

Kane, M.T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

Keenan, J. & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47*, 125-135.

Lafuente-Martínez, M., Lévêque, O., Benítez, I., Hardebolle, C., & Dehler Zufferey, J. (2022). Assessing computational thinking: Development and validation of the Algorithmic Thinking Test for adults. *Journal of Educational Computing Research, 60*(6), 1436-1463. https://doi.org/10.1177/07356331211057819

Lane, S. (2014). Validity evidence based on testing consequences. *Psicothema, 26*(1), 127-135. https://doi.org/10.7334/psicothema2013.258

Lee, S., & Winke, P. (2018). Young learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35*(2), 239-269. https://doi.org/10.1177/0265532217704009

Luesia, J. F., Sánchez-Martín, M., & Benítez, I. (2021). The effect of personal values on academic achievement. *Psychological Test and Assessment Modeling, 63*(2), 168-190.

Marchant, G. J. & Paulson, S. E. (2005, January 21). The relationship of high school graduation exams to graduation rates and SAT scores. *Education Policy Analysis Archives, 13*(6). http://epaa.asu.edu/epaa/v13n6/

Martone, A., & Sireci, S. G. (2009). Evaluating alignment between curriculum, assessment, and instruction. *Review of Educational Research, 79*(4), 1332-1361. https://doi.org/10.3102/0034654309341375

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement,* (3rd ed., pp. 13-100). American Council on Education.

Mislevy, R. J. (2019). On integrating psychometrics and learning analytics in complex assessments. In H. Jiao, R. W. Lissitz, & A. van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 1–52). Information Age.

Morris, S. B., & Dunleavy, D. M. (2016). *Adverse impact analysis: Understanding data, impact, and risk*. Routledge.

Muñiz, J., & Fonseca-Pedrero, E. (2019). Ten steps for test development. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

Newman, D. A., Tang, C., Song, Q. C., & Wee, S. (2022). Dropping the GRE, keeping the GRE, or using GRE-optional admissions? Considering tradeoffs and fairness. *International Journal of Testing, 22*(1), 43-71. https://doi.org/10.1080/15305058.2021.2019750

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*, 301–319. https://doi.org/10.1037/a0032969

Noble, T., Rosebery, A., Suarez, C., Warren, B., & O'Connor, M. C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education, 27*(4), 248–260. https://doi.org/10.1080/08957347.2014.944309

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136-144. https://doi.org/10.7334/psicothema2013.259

Randall, J. (2021). "Color-neutral" is not a thing: Redefining construct definition and representation through a justice-oriented critical antiracist lens. *Educational Measurement: Issues and Practice, 40*(4), 82-90.

Reynolds, K. A. & Moncaleano, S. (2021). Digital module 26: Content alignment in standards-based educational assessment. *Educational Measurement: Issues & Practice, 40*(3), 127-128.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothema, 26*(1), 108-116. https://doi.org/10.7334/psicothema2013.260

Rushton, P. W., Routhier, F., Miller, W. C., Auger, C., & Lavoie, M. P. (2015). French-Canadian translation of the WheelCon-M (WheelCon-MF) and evaluation of its validity evidence using telephone administration. *Disability and Rehabilitation, 37*(9), 812-819. https://doi.org/10.3109/09638288.2014.941019

Russell, M. (2022). Clarifying the terminology of validity and the investigative tages of validation. *Educational Measurement: Issues and Practice, 41*(2), 25-35.

Segool, N. K., Carlson, J. S., Goforth, A. N., von der Embse, N. & Barterian, J. N. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*, 489-499.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Sinha, R., Oswald, F., Imus, A & Schmitt, N. (2011). Criterion-focused approach to reducing adverse impact in college admissions. *Applied Measurement in Education, 24*, 137-161, https://doi.org/10.1080/08957347.2011.554605

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research, 45*(1), 83-117.

Sireci, S. G. (2016a). Comments on valid (and invalid?) commentaries. *Assessment in Education: Principles, Policy & Practice, 23*, 319-321. http://dx.doi.org/10.1080/0969594X.2016.1158694

Sireci, S. G. (2016b). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*, 226-235. https://doi.org/10.1080/0969594X.2015.1072084

Sireci, S. G. (2020). De-"constructing" test validation. Chinese/English *Journal of Educational Measurement and Evaluation*| 教育测量与评估双语季刊*, 1*(1), Article 3.

Sireci, S., & Faulkner-Bond, M. (2014). Validity evidence based on test content. *Psicothema, 26*(1), 100-107. https://doi.org/10.7334/psicothema2013.256

Sireci, S. G., & Geisinger, K. F. (1998). Equity issues in employment testing. In J.H. Sandoval, C. Frisby, K. F. Geisinger, J. Scheuneman, & J. Ramos-Grenier (Eds.), *Test interpretation and diversity* (pp. 105-140). American Psychological Association.

Sireci, S. G., & Greiff, S. (2019). On the importance of educational tests. *European Journal of Psychological Assessment, 35*, 297-300. https://doi.org/10.1027/1015-5759/a000549.

Sireci, S. G., Lim, H., Rodriguez, G., Banda, E., & Zenisky, A. (2018, April 12-16). *Evaluating criteria for validity evidence based on test content* [Conference presentation]. Annual meeting of the National Council on Measurement in Education, New York, United States.

U.S. Equal Employment Opportunity Commission (2010). Fact sheet on employment tests and selection procedures. Washington, DC: Author. Available at https://urldefense.com/v3/__https://www.eeoc.gov/policy/docs/factemployment_procedures.html__;!!D9dNQwwGXtA!UIxPPhFeyMiN5JZpTwpvh_T8FQTdj7TaEssb4sMT8hiLFhN1ssQa2qSwdQ1SbkMr58y5-dBdeGBiWt4DAts$

Whitney, C. R., & Candelaria, C. A. (2017). The effects of No Child Left Behind on children's socioemotional outcomes. *AERA Open, 3*(3), 1-21. https://doi.org/10.1177/2332858417726324

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O'Donnell, F., Wells, C. S., Padellaro, F., Jung, H., Banda, E., Pham, D. Hong, S., Park, Y., Botha, S., Lee, M, & Garcia, A. (2018, September). Massachusetts Adult Proficiency Tests for college and career readiness: Technical manual. *Center for Educational Assessment research report No. 974*. Center for Educational Assessment.